arXiv:cond-mat/0402499v2 [cond-mat.dis-nn] 20 Feb 2004

Detecting communities in large networks

A. Capocci[1], V.D.P. Servedio[12], G. Caldarelli[12], F. Colaiori[2]
[1]*Centro Studi e Ricerche e Museo della Fisica "E. Fermi", Compendio Viminale, Roma, Italy*
[2]*INFM UdR Roma1-Dipartimento di Fisica Università "La Sapienza", P.le A. Moro 5, 00185 Roma, Italy*

We develop an algorithm to detect community structure in complex networks. The algorithm is based on spectral methods and takes into account weights and links orientations. Since the method detects efficiently clustered nodes in large networks even when these are not sharply partitioned, it turns to be specially suitable to the analysis of social and information networks. We test the algorithm on a large-scale data-set from a psychological experiment of word association. In this case, it proves to be successful both in clustering words, and in uncovering mental association patterns.

PACS numbers: : 89.75.Hc, 89.75.Da, 89.75.Fb

Measurements and exact results concerning the clustering patterns of networks mainly concern the occurrence of regular motifs [1, 2, 3, 4] and their correlations [5, 6, 7]. However, many social and information networks, such as the World Wide Web, turn out to be approximately partitioned into communities of irregular shape: for example, web pages focusing on similar topics are strongly mutually connected and have a weaker linkage to the rest of the Web. The design of methods to partition a graph into several meaningful highly interconnected components have then become a compelling application of graph theory to biological, social and information networks [8, 9, 10, 11].

Detecting the community structure in information networks allows one to mine information in a more efficient way, narrowing the exploration of a network as large as the World Wide Web (about 10^9 nodes) to a limited portion of it. When used in the analysis of large collaboration networks, such as company or universities, communities reveal the informal organization and the nature of information flows through the whole system [12, 13].

There are several empirical methods to detect communities. The most successful algorithm, recently introduced by [10] (NG–algorithm), is based on the edge betweenness, that measures the fraction of all shortest paths passing on a given link, or, alternatively, the probability that a random walk on the network runs over that link. By removing links with high betweenness, one progressively splits the whole network into disconnected components, until the network is decomposed in communities consisting of one single node. The outcome of the algorithm is represented by a dendrogram, i.e. a tree–like diagram where each branching corresponds to a splitting event. Though this method has been shown to be very powerful in cases where some a priori knowledge of the a community structure is given, it has two main disadvantages: first, that it does not give an indication of the resolution of the clustering, and thus it needs extra information as input (like the expected number of clusters); second, that its outcome is independent on how sharp the partitioning of the graph is. In the same spirit, [14] proposed an algorithm based on local analogues of the edge betweenness. This has the advantage of being faster, but has the same drawbacks on the NG–algorithm.

An alternative way to tackle the problem, which is the one we pursue, is by spectral analysis. Previous approaches to graph partitioning from spectral analysis have been mostly developed in the computer science community to the purpose of finding the best allocation of processes on processors in parallel computers, and are based on iterative bisection. When applied to find communities structures these methods have the disadvantage that repeated bisection is not guaranteed to reach the best or most natural partition in general cases. Moreover, they suffer from the same limitation of the algorithm based on the edge betweenness, since they give no indication of when the bisection should terminate, and thus need extra information on the expected number of communities.

Our aim in this paper is to develop some spectral based algorithm able to reveal the structure of a complex network, which could be blurred by the bias artificially overimposed by the iterative bisection constraint. Such a method should be able to conjugate the power of spectral analysis to the caution needed to reveal an underlying structure when there is no clear cut partitioning, as is often the case in real networks.

Spectral methods are based on the analysis of the adjacency matrix A [15, 16, 17], whose element a_{ij} is equal to 1 if i points to j and 0 otherwise. In particular, such methods analyze simple functions of A: the Laplacian matrix $L = K - A$ and the Normal matrix $N = K^{-1}A$, where K is the diagonal matrix with elements $k_{ii} = \sum_{j=1}^{S} a_{ij}$ and S is the number of nodes in the network. In most approaches, referring to undirected networks, A is assumed to be symmetric.

The matrix N has always the largest eigenvalue equal to one, associated to a trivial constant eigenvector, due to row normalization. In a network with an apparent cluster structure, N has also a certain number $m-1$ of eigenvalues close to one, where m is the number of well defined communities, the remaining eigenvalues lying a gap away from one. The eigenvectors associated to these first $m-1$ nontrivial eigenvalues, also have a characteristic structure: the components corresponding to nodes within the same cluster have very similar values x_i, so that, as long as the partition is sufficiently sharp, the profile of each eigenvector, sorted by components, is step–like. The number of steps in the profile corresponds again the number m of communities. A similar information is encoded in the non-negative definite Laplacian matrix, where the eigenvalues close to zero are associated to clusters.

The study of the eigenvectors profiles and the eigenvalues has practical use only when a clear partition exists, which is rarely the case. In most common occurrences, the number of nodes is too large and the separation between the different communities is rather smooth. Thus communities cannot be simply detected by looking at the first nontrivial eigenvector. We resolve this issue by combining information from the first few eigenvectors, and extracting the community structure from correlations between the same components in different eigenvectors.

To describe the method in detail and understand why it works, it is instructive to recast the eigenproblem into an optimization problem. With the most general applications in mind, instead of the adjacency matrix A, we focus on the weight matrix W, whose elements w_{ij} are assigned the intensity of the link (i, j). We consider undirected graphs first, and then we pass to the most general directed case. Consider the following constrained optimization problem: Let $z(\mathbf{x})$ be defined as

$$z(\mathbf{x}) = \frac{1}{2} \sum_{i,j=1}^{S} (x_i - x_j)^2 w_{ij} \,, \qquad (1)$$

where x_i are values assigned to the nodes, with some constraint on the vector $\mathbf{x}$, expressed by

$$\sum_{i,j=1}^{S} x_i x_j m_{ij} = 1 \,, \qquad (2)$$

where m_{ij} are elements of a given symmetric matrix M.

The stationary points of z over all $\mathbf{x}$ subject to the constraint (2) are the solutions of

$$(D - W)\mathbf{x} = \mu M \mathbf{x} \,, \qquad (3)$$

where D is the diagonal matrix $d_{ij} = \delta_{ij} \sum_{k=1}^{S} w_{ik}$, and μ is a Lagrange multiplier.

Different choices of the constraint M leads to different eigenvalues problems: for example choosing $M = D$ leads the eigenvalues problem $D^{-1}W\mathbf{x} = (1-2\mu)\mathbf{x}$, while $M = 1$ leads to $(D - W)\mathbf{x} = \mu\mathbf{x}$. Thus $M = D$ and $M = 1$, corresponds to the eigenproblems for the (generalized) Normal and Laplacian matrix respectively.

Thus, solving the eigenproblem is equivalent to minimizing the function (1) with the constraint (2), were the x_i's are eigenvectors components. The absolute minimum corresponds to the trivial eigenvector, which is constant. The other stationary points correspond to eigenvectors where components associated to well connected nodes assume similar values.

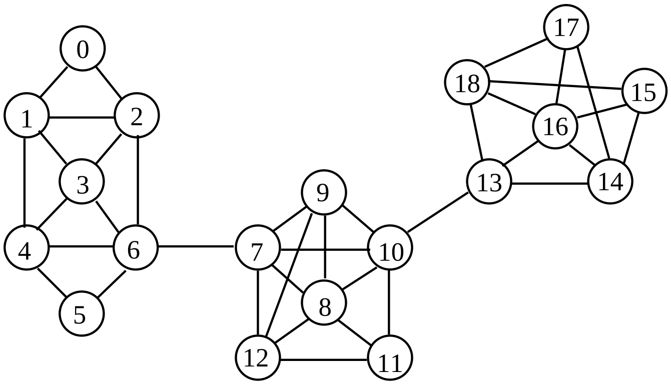

FIG. 1: Network employed as an example, with $S = 19$ and random weights between 1 and 10 assigned to the links. Three clear clusters appear, composed by nodes $0 - 6$, $7 - 12$ and $13 - 19$.

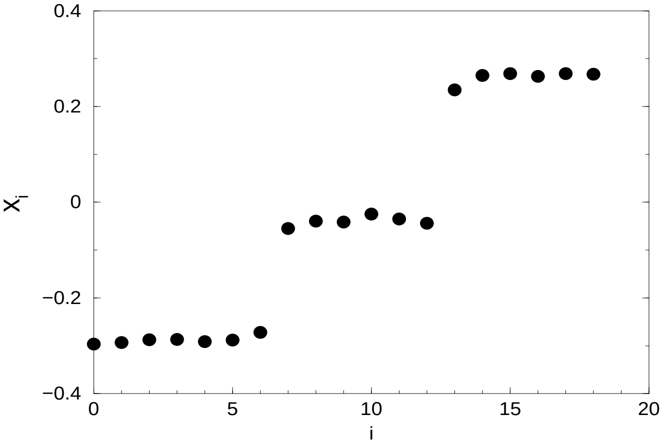

FIG. 2: Values of the 2nd eigenvector components for matrix $D^{-1}W$ relative to the graph depicted in figure 1.

In order to compute cluster sizes and distribution, methods such as bisection or edge-betweenness based ones are very poor in detect the end of the recursive splitting. Our approach, instead, immediately detects the number of clear clusters from the eigenvectors profile.

As an illustrative example, we show in Fig.2 the profile of the second eigenvectors of $D^{-1}W$ corresponding to the simple graph shown in Fig.1 with $S = 19$ nodes, where random weights between 1 and 10 were assigned to the links. The components of the eigenvectors assume approximatively constant values on nodes belonging to the same community. Thus, the number of communities emerges naturally and it is not needed as input, .

However, as aforementioned, when dealing with large networks with no clear partitioning, the precise value of the eigenvector components is of little use. In such situations, the typical eigenvector profile is not step-like, but resembles a continuous curve. Nevertheless, our method can still be applied, and efficiently detects sets of well connected nodes. In fact, components corresponding to nodes belonging to the same communities are still strongly correlated taking, in each eigenvector, similar values among themselves. Thus, a natural way to identify communities in an automatic manner, is by measuring the correlation

$$r_{ij} = \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{[(\langle x_i^2 \rangle - \langle x_i \rangle^2)(\langle x_j^2 \rangle - \langle x_j \rangle^2)]^{\frac{1}{2}}}\,, \tag{4}$$

where the average $\langle \cdot \rangle$ is over the first few nontrivial eigenvectors. The quantity r_{ij} measures the community closeness between node i and j. Though the performance may be improved by averaging over more and more eigenvectors, with increased computational effort, we find that indeed a small number of eigenvectors suffices to identify the community to which nodes belong, even in large networks.

When dealing with a directed network, links do not correspond to any equivalence relation. Rather, pointing to common neighbors is a significant relation, as suggested in the sociologists' literature where this quantity measures the so-called *structural equivalence* of nodes [18]. Accordingly, in a directed network, clusters should be composed by nodes pointing to a high number of common neighbors, no matter their direct linkage. For directed networks, we thus modify our method in the streamline of the HITS algorithm [17]. The HITS algorithm was proposed on empirical bases to find the main communities in large oriented networks. It assumes that the largest components (in the absolute value) of eigenvectors of the matrices AA^T and A^TA correspond to highly clustered nodes belonging to a single community. Such algorithm efficiently detects the main communities, even when these are not sharply defined. However, it becomes computationally heavy when one is interested in minor communities, which correspond to smaller eigenvalues. As explained in the undirected case, we tackle this issue by combining information from the first few eigenvectors, and extracting the community structure from correlations between the same components in different eigenvectors.

To detect the community structure in a directed network, we therefore replace, in the previous analysis, the matrix W with a matrix $Y = WW^T$. This corresponds to replacing the directed network with an undirected weighted network, where nodes pointing to common neighbors are connected by a link, whose intensity is proportional to the total sum of the weights of the links pointing from the two original nodes to the common neighbors. Then, one performs the analysis on the undirected network as described previously. Thus, the function to minimize in this case is

$$y(\mathbf{x}) = \sum_{ijl}^{1,S} (x_i - x_j)^2 w_{il} w_{jl}\,. \tag{5}$$

Defining Q as the diagonal matrix $q_{ij} = \delta_{ij} \sum_{lj=1}^{S} w_{il} w_{jl}$, the eigenvalue problem for the analogous of the generalized normal matrix,

$$Q^{-1} Y \mathbf{x} = \lambda \mathbf{x} \tag{6}$$

is equivalent to minimizing the function (5) under the constraint $\sum_{ijl=1}^{S} x_i x_j q_{ij} = 1$.

Tested on simple examples of directed networks, the algorithm associated to the minimization of y, outperforms the one based on the minimization of z.

To test this spectral correlation-based community detection method on a real complex network, we apply the algorithm to data from a psychological experiment reported in reference [19]. Volunteering participants to the research had to respond quickly by freely associating a word (response) to another word given as input (stimulus), extracted by a fixed subset. Scientists conducting the research have recorded all the stimuli and the associated responses, along with the occurrence of each association. In the same spirit of past works [20], we construct a network were words are nodes, and directed links are drawn from each stimulus to the corresponding responses, assuming that a link is oriented from the stimulus to the response. The resulting network includes $S = 10616$ nodes, with an average in-degree equal to about 7. Taking into account the frequency of responses to a given stimulus, we construct the weighted adjacency matrix W. In this case, passing to the matrix Y means that we expect stimuli giving rise to the same response to be correlated.

The large-scale properties of semantic [19, 21, 22] and syntactic networks[23] corresponding to different languages have been examined in past literature, mainly based on dictionaries and texts: a strong similarity has emerged in such surveys, showing that statistical features must refer to a common underlying structure rather than to individual cultures. Interestingly, word graphs studied so far are found to be complex networks, characterized by the small world property and by power-law degree distribution independently of the specific definition of the network [24].

The word association network is an ideal playground to test our algorithm as, despite the large size of the networks, the quality of clustering can be evaluated by a direct inspection to the yieldings. In large databases like this, were a partition in communities is not defined in a natural manner, there is no definite answer to what the best partition is. Rather, one is interested in finding groups of highly correlated nodes, or groups of nodes highly connected to a given one. Table I shows the most correlated words to three test-words. The correlation are computed by averaging over just 10 eigenvectors of the matrix $Q^{-1}Y$: the results appear to be quite satisfactory, already with this small number of eigenvectors.

Besides the performance in finding clusters of correlated words, our results are suggestive of the criteria according to which the participants to the experiment have associated words. As we observed, free associations are

science	1	literature	1	piano	1
scientific	0.994	dictionary	0.994	cello	0.993
chemistry	0.990	editorial	0.990	fiddle	0.992
physics	0.988	synopsis	0.988	viola	0.990
concentrate	0.973	words	0.987	banjo	0.988
thinking	0.973	grammar	0.986	saxophone	0.985
test	0.973	adjective	0.983	director	0.984
lab	0.969	chapter	0.982	violin	0.983
brain	0.965	prose	0.979	clarinet	0.983
equation	0.963	topic	0.976	oboe	0.983
examine	0.962	English	0.975	theater	0.982

TABLE I: The words most correlated to *science*, *literature* and *piano* in the eigenvectors of $Q^{-1}WW^T$. Values indicate the correlation.

made by synonymy or antinomy, syntactic role, and even by analogous sensory perception.

In conclusion, we have introduced a new method to detect communities of highly connected nodes within a network. The method is based on spectral analysis and takes into account the presence of weighted links between nodes. Unlike previous spectral approaches, our method is not based on iterative bisection. We have tested our algorithm on a real network instance, built upon the records of a psychological experiments. The algorithm proves to be successful in clustering nodes (in this case, words) according to reasonable criteria, and provides an automatic way to extract the most connected sets of nodes to a given one in a set of over 10^4. Given the broad range of applicability, such method suggests a reliable way of clustering large-scale networks occurring in different fields, including biology, computer science and sociology.

We enjoyed useful discussion and suggestions by Ramon Ferrer i Cancho and Miguel-Angel Muñoz.

We acknowledge partial support from the FET Open Project IST-2001-33555 COSIN.

[1] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).

[2] S. N. Dorogovtsev and J. F. F. Mendes, *Adv. in Phys.* **51**, 1079 (2002).

[3] J.P. Eckmann, E. Moses, *PNAS* **99** (9), 5825 (2002).

[4] G. Bianconi and A. Capocci, *Phys. Rev. Lett.* **90** ,078701 (2003).

[5] G. Caldarelli, R. Pastor-Satorras and A. Vespignani, cond-mat/0212026 (2002).

[6] G. Caldarelli, A. Capocci, P. De Los Rios, M.A. Muñoz, *Phys. Rev. Lett.* **89** 258702 (2002).

[7] A. Capocci, G. Caldarelli, P. De Los Rios, *Phys. Rev.* E **68** 047101 (2003).

[8] I. Simonsen, K. A. Eriksen, S. Maslov, K. Sneppen, cond-mat/0312476 (2003), to appear in *Physica A*

[9] S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, *The VLDB Journal*, 639 (1999).

[10] M. Girvan and M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **99**, 8271 (2002).

[11] M. E. J. Newman, *SIAM Review* **45**, 167 (2003).

[12] B. Huberman, J. Tyler and D. Wilkinson, in *Communities and technologies*, M. Huysman, E. Wegner and V. Wulf, eds. Kluwer Academic (2003).

[13] R. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt and A. Arenas, *Phys. Rev. E* **68** 065103 (2003)

[14] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, submitted for publication, preprint cond-mat/0309488

[15] K. M. Hall, *Management Science* **17**, 219 (1970).

[16] A.J. Seary and W.D. Richards, W.D. *Proceedings of the International Conference on Social Networks Volume 1: Methodology*, 47 (1995).

[17] J. Kleinberg, *Journal of the ACM* **46** (5) 604 (1999).

[18] M. E. J. Newman, *Eur. Phys. J. B*, in press.

[19] M. Steyvers, J. B. Tenenbaum, preprint cond-mat/0110012, submitted for publication.

[20] L. Da Fontoura Costa, preprint cond-mat/0309266, submitted for publication

[21] M. Sigman and G. A. Cecchi, *Proc. Natl. Acad. Sci. USA* **99**, 1742 (2002).

[22] S.N. Dorogovtsev and J.F.F. Mendes,*Proc. Royal Soc. London B* **268**, 2603 (2001)

[23] R. Ferrer i Cancho, R. V. Solé and R. Köhler, Santa Fe Institute Working Paper n. 03-06-042 (2003), submitted for publication.

[24] R. Ferrer i Cancho, R.V. Solé, *Proc. R. Soc. Lond.* B **268** 2261 (2001).

[25] M. E. J. Newman, *Phys. Rev. Lett.* **89**, 208701 (2002).

[26] A. Vázquez, R. Pastor-Satorras and A. Vespignani, *Phys. Rev. E* **65**, 066130 (2002).

[27] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).

[28] G.R. Kiss, C. Armstrong, R. Milroy and J. Piper, *An associative thesaurus of English and its computer analysis.* In A.J. Aitken, R.W. Bailey and N. Hamilton-Smith, (Eds.), *The Computer and Literary Studies*, Edinburgh University Press (1973).